Supplementary Information

Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs

CONTENTS

Supplementary Table 1. Case-control studies comprising PGC-SCZ

Sub-set	Sample Cohort	Genotyping Platform	In PGC-SCZ GWAS Cases	In PGC-SCZ GWAS Controls	Used here Cases	Used here Controls
ISC	ISC-Aberdeen	Affymetrix 5.0	720	698	696	661
ISC	ISC-Cardiff	Affymetrix 6.0	527	609	518	597
ISC	ISC-Dublin	Affymetrix 6.0	270	860	264	838
ISC	ISC-Edinburgh	Affymetrix 6.0	368	284	360	279
ISC	ISC-London	Affymetrix 5.0 & 500K	518	491	516	491
ISC	ISC-Portugal	Affymetrix 5.0	346	215	324	189
ISC	ISC-SW1	Affymetrix 5.0	168	167	164	166
ISC	ISC-SW2	Affymetrix 6.0	390	229	378	224
ISC	Total		**3307**	**3553**	**3220**	**3445**
MGS	MGS	Affymetrix 6.0	**2679**	**2484**	**2571**	**2419**
OTH	SGENE-Bonn	Illumina 550K	474	1,304	475	1297
OTH	SGENE-CH	Illumina Human 610-Quad	482	457	476	450
OTH	SGENE-MUN	Illumina 300K	434	351	424	351
OTH	SGENE-TOP3	Affymetrix 6.0	248	351	244	348
OTH	SGENE-UCLA	Illumina 550K	704	631	695	623
OTH	Cardiff	Affymetrix 550K	472	2,934	465	2917
OTH	CATIE	Affymetrix 500K; Perlegen 164K	402	207	379	209
OTH	Zucker Hillside	Affymetrix 500K	192	190	138	112
OTH	Total		**3408**	**6425**	**3296**	**6307**
Total	PGC-SCZ		**9394**	**12462**	**9087**	**12171**

The case-control samples contributing to PGC-SCZ are described in detail in the Supplementary Material of Ripke *et al*[1]. All PGC-SCZ data sets underwent stringent quality control (QC) separately prior to imputation but using the same common parameters:

- Missing rate per SNP <0.05 (prior to sample removal below)
- Missing rate per individual <0.02
- Missing rate per SNP <0.02 (after sample removal above)
- Missing rate per SNP difference cases-controls <0.02
- SNP frequency difference to HapMap <0.15
- Hardy-Weinberg Equilibrium (H-W) test (controls) $p<10^{-6}$

The reduced numbers used here compared to the PGC-SCZ GWAS reflects more stringent selection of individuals who are unrelated in the classical sense, average relatioship (equation 2) < 0.05.

Supplementary Table 2. Pilot analyses using the ISC and MGS GWAS data; estimates of proportion of variance in liability explained by SNPs (h^2) **See Supplementary Note**

Analysis	Source	Genotypes	Cases	Controls	SNPs	LR	h^2 (s.e.)	h^2 (s.e.) Adjusted
ISC								
1	Baseline stringent QC	G	2,538	2,788	287,782	141	0.33 (0.03)	0.38 (0.03)
2	Extreme QC	G	2,538	2,788	196,322	100	0.27 (0.03)	0.32 (0.03)
3	In PGC-SCZ	I	3,220	3,445	915,354	165	0.28 (0.02)	0.30 (0.02)
4	In PGC-SCZ	G	3,220	3,445	192,477	146	0.26(0.02)	0.30 (0.02)
5	In PGC-SCZ ex London	I	2,704	2,954	915,354	138	0.31 (0.03)	0.33 (0.03)
MGS								
6	Baseline stringent QC	G	1,294	1,862	590,622	41	0.36 (0.06)	0.38(0.06)
7	Extreme QC	G	1,294	1,862	491,851	31	0.30(0.05)	0.32(0.06)
8	In PGC-SCZ	I	2,571	2,419	915,354	135	0.34 (0.03)	0.35 (0.03)
9	In PGC-SCZ	G	2,571	2,419	460,939	120	0.32 (0.03)	0.35 (0.03)

G: genotyped SNPs only; I imputed SNPs: LR: likelihood ratio test statistic; h^2 Adjusted estimate of proportion of variance in liability to schizophrenia explained by SNPs adjusted to account for incomplete linkage disequilibrium between causal variants and SNPs; SE standard error of h^2. NB All analyses include 4 ancestry principal components. In main text 20 ancestry principal components are used.

Supplementary Table 3. Estimated proportion of variance in liability attributable to SNPs on the X chromosome (h^2) **See Supplementary Note**

a) Bivariate analyses in which trait 1 comprises only male data and trait 2 comprises only female data

	Male h^2 (s.e.)	Female h^2 (s.e.)	r
PGC-SCZ Autosome	0.244 (0.013)	0.253 (0.019)	0.894 (0.058)
MGS-ISC Autosome	0.268 (0.032)	0.253 (0.045)	0.902 (0.139)
MGS-ISC X	0.008 (0.006)	0.010 (0.008)	0.909 (0.688)

ISC and MGS data sets combined, comprised 21380 genotyped SNPs on the X chromosome

b) Comparing equal variance vs. dosage compensation models

	LR	h^2 (s.e.)
Equal variance model (male variance = female variance)	7.64	0.009 (0.004)
dosage compensation model- female variance (male variance=2*female variance)	11.11	0.007 (0.002)

Supplementary Table 4. Estimated proportion of variance in liability (h^2) to schizophrenia explained by SNPs in CNS+ genes, other genes and not in genes.

	No. Genes	No. SNPs	~Mb	Mb as % of total	h^2 (s.e.)	h^2 as % of total (s.e.)
CNS+ genes	2725	195044	547	20[†]	0.07 (0.01)	31 (2)[†]
Other genes	14804	355562	1069	39	0.08 (0.01)	35 (2)
Not in genes		364748	1155	42	0.08 (0.01)	34 (2)
Total			2772	100	0.23	100

Estimates based on 915354 imputed SNPs; SE standard error of h^2. † The proportion 31 % (s.e. 2%) is significantly different to the proportion 20% p =7.6 x 10^{-8}.

Supplementary Table 5. Estimated proportion of variance in liability to schizophrenia explained by SNPs, partitioned by minor allele frequency (MAF), for a single analysis of all five MAF categories (Joint) or five individual analyses of each MAF bin.

MAF of SNPs	Number of SNPs	h^2 (s.e.) ISC (Joint)	MGS (Joint)	OTH (Joint)	PGC-SCZ (Joint)	PGC-SCZ (Separate)	PGC-SCZ (Proportion[a])
<0.1	158497	0.03 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.01)	0.06 (0.01)	0.26 (0.04)
0.1 – 0.2	209466	0.05 (0.02)	0.05 (0.03)	0.07 (0.02)	0.05 (0.01)	0.12 (0.01)	0.52 (0.05)
0.2 – 0.3	191270	0.05 (0.02)	0.08 (0.03)	0.06 (0.02)	0.04 (0.01)	0.13 (0.01)	0.57 (0.05)
0.3 – 0.4	181782	0.07 (0.02)	0.12 (0.03)	0.06 (0.01)	0.06 (0.01)	0.13 (0.01)	0.57 (0.05)
0.4<	174339	0.07 (0.02)	0.03 (0.02)	0.05 (0.01)	0.05 (0.01)	0.11 (0.01)	0.48 (0.04)
sum	915354	0.27	0.30	0.26	0.22		

a. Proportion of the variance explained by all SNP (0.22) explained by SNPs in each MAF bin from the separate analyses

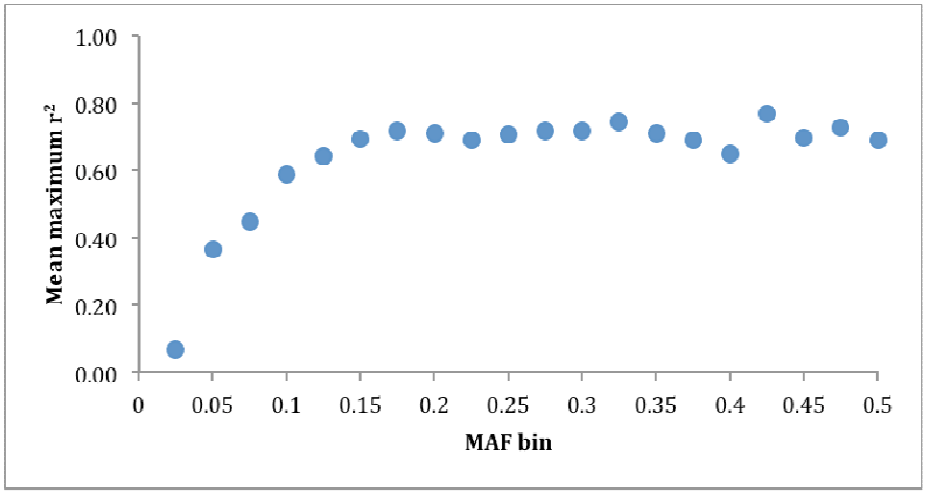
Supplementary Table 6. Estimated proportion of variance explained (h^2) using simulated data based on all PGC-SCZ imputed genotypes. **See Supplementary Note**

	All variance in liability associated with SNPs with MAF < 0.1				All variance in liability associated with SNPs across the MAF spectrum	
	Joint analysis	Joint analysis	Separate analysis	Separate proportion[a]	Joint analysis	Separate analysis
25% of variance in liability attributed to SNPs with MAF < 0.1						
< 0.1	0.25 (0.004)	Not included	0.25 (0.004)	1.00 (0.016)	0.05 (0.005)	0.08 (0.004)
0.1 - 0.2	0.00 (0.004)	0.06 (0.004)	0.07 (0.003)	0.28 (0.012)	0.05 (0.004)	0.12 (0.004)
0.2 - 0.3	0.00 (0.003)	0.01 (0.003)	0.04 (0.002)	0.16 (0.008)	0.05 (0.004)	0.13 (0.003)
0.3 - 0.4	0.00 (0.002)	0.01 (0.003)	0.03 (0.002)	0.12 (0.008)	0.05 (0.003)	0.12 (0.002)
0.4 –0.5	0.00 (0.002)	0.00 (0.003)	0.02 (0.002)	0.08 (0.008)	0.05 (0.002)	0.11 (0.002)
Total	0.25	0.08			0.25	
50% of variance in liability attributed to SNPs with MAF < 0.1						
< 0.1	0.50 (0.003)	Not included	0.50 (0.003)	1.00 (0.007)	0.08 (0.003)	0.16 (0.003)
0.1 - 0.2	0.00 (0.002)	0.13 (0.004)	0.14 (0.003)	0.28 (0.007)	0.12 (0.004)	0.26 (0.005)
0.2 - 0.3	0.00 (0.003)	0.02 (0.004)	0.08 (0.002)	0.16 (0.004)	0.10 (0.003)	0.26 (0.004)
0.3 - 0.4	0.00 (0.003)	0.01 (0.004)	0.06 (0.003)	0.12 (0.006)	0.10 (0.004)	0.25 (0.004)
0.4 –0.5	0.00 (0.003)	0.02 (0.003)	0.05 (0.003)	0.10 (0.006)	0.10 (0.004)	0.22 (0.004)
Total	0.50	0.18			0.50	
80% of variance in liability attributed to SNPs with MAF < 0.1						
< 0.1	0.80 (0.003)	Not included	0.80 (0.002)	1.00 (0.003)	0.14 (0.003)	0.28 (0.004)
0.1 - 0.2	0.00 (0.002)	0.21 (0.004)	0.24 (0.004)	0.30 (0.005)	0.18 (0.004)	0.42 (0.004)
0.2 - 0.3	0.00 (0.002)	0.05(0.003)	0.14 (0.003)	0.17 (0.004)	0.17 (0.004)	0.44 (0.002)
0.3 - 0.4	0.00 (0.002)	0.02 (0.004)	0.10 (0.004)	0.13 (0.005)	0.16 (0.005)	0.42 (0.003)
0.4 –0.5	0.00 (0.002)	0.02 (0.003)	0.09 (0.004)	0.11 (0.005)	0.15 (0.004)	0.37 (0.003)
Total	0.80	0.30			0.80	

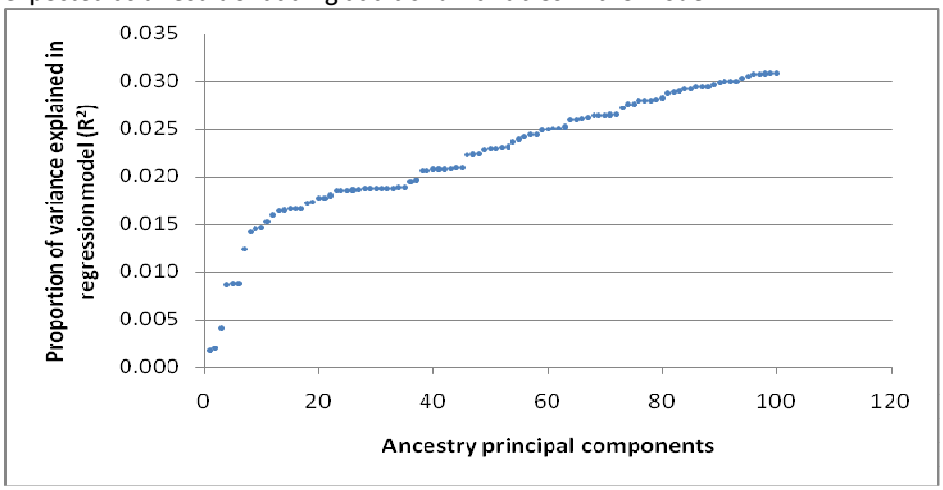
a.Proportion of the variance explained by all SNP (ie 0.25, 0.5 or 0.8) explained by SNPs in each MAF bin from the separate analyses. Values in parenthesis are standard errors of the mean estimates over 10 replicates.

SUPPLENTARY FIGURES.

Supplementary Figure 1. Mean maximum linkage disequilibrium r^2 measure between a random sample of SNPs in 1000G and SNPs on the Affymetrix 6.0 genotyping chip. We used genotype data from the 1000 genomes (1000G) project[2] to provide an estimate of the proportion of known variance in the genome explained by SNPs genotyped in the PGC-SCZ. We selected 10,000 autosomal SNPs ("target" SNPs) at random from the 37 million SNPs in European 1000G (381 samples); this random selection ensures that these target SNPs are unlikely to be in LD with each other. After excluding monomorphic SNPs, 4292 SNPs were assigned to MAF bins; 52% of these had frequency in the first MAF bin, defined as MAF < 0.025. We calculated the maximum LD r^2 between each target SNP and SNPs within ±1Mb about it included in the MGS genotyped SNP set (Affymetrix 6.0 ~460k SNPs) using European 1000G data. These results are consistent with those published using the pilot 1000G data[3].



Supplementary Figure 2. Proportion of variance explained from regression of case-control status on increasing numbers of ancestry principal components. Case control status is the dependent variable and sex and sample cohort were included as covariates. The principal components are entered in their eigenvalue order. The increase in R^2 with > 20 principal components is linear, an increase expected as a result of adding additional variables in the model.



SUPPLEMENTARY NOTE.

1) Notes on Supplementary Table 2

The purpose of Supplementary Table 2 is to demonstrate that the QC applied to the PGC-SCZ data did not introduce any bias which would inflate the estimates of variance explained by SNPs. To do this we made comparisons with our own QC and extensive original analyses of the ISC and MGS data sets. Those original analyses utilised plate/batch effects, information which was not available for the full PGC-SCZ data; our analyses confirmed that the PGC-SCZ QC procedure is unlikely to introduce inflated estimates of variance explained by SNPs.

ISC

Our preliminary analyses first used only genotyped SNPs from the ISC GWAS data set provided to us directly by the ISC, comprising 3,322 European individuals with schizophrenia and 3,587 controls (ancestry outliers had been excluded)[4]. To check that the estimated h^2 attributed to SNPs could not reflect systematic bias caused from batch effects or genotyping errors we investigated the impact of imposing extremely stringent QC thresholds. We excluded the ISC-London sample because cases and controls were genotyped on different platforms, for all other cohorts cases and controls were genotyped together. We also excluded subjects so that no pair of individuals had a genome-wide similarity coefficient (equation 2) > 0.05. We included genotyping plate as an additional random effect in the linear mixed model. The baseline stringent QC excluded SNPs if they had

- MAF < 0.01
- Missingness rate (MR) > 0.05
- H-W test p<0.0001

We compared the estimates h^2 to those obtained with extreme QC , defined as exclusion of SNPs with p<0.05 for

- the H-W test, including in any individual data set contributing to the for ISC
- the test for differential missingness of genotypes between cases and controls,
- two-locus QC test[5] applied in a very conservative way so that each SNP was explicitly paired and tested with each of 20 flanking markers.

The numbers of SNPs, cases and controls remaining after these QC steps are listed in the table. Comparison of the estimates of h^2 between analyses 1 and 2 shows a lower estimate, but not significantly so, when extreme QC is applied, implying that the variance detected is unlikely to be explained by genotyping artefacts. The extreme QC leads to a loss of SNPs so the reduced estimate of h^2 may reflect reduced LD between the genotyped SNPs and causal variants. In order to make fairer comparisons between alternatives because of the differing numbers of SNPs we adjusted the estimated variance to account for the imperfect LD between the genotyped SNPs and causal variants. The adjustment assumes that causal variants have allele frequency spectrum and LD patterns with genotyped SNPs similar to the allele frequency spectrum of, and LD patterns between, genotyped SNPs[6,7] used in the analysis. If the SNPs remaining in the extreme QC set were a random set of the SNPs in the baseline QC set, then the adjusted h^2 estimated from the two analyses would be the same. However, we noted that the extreme QC preferentially excluded low minor allele frequency (MAF) SNPs and so the variance captured by these SNPs could not be recovered by the adjustment.

In order to use the PGC-SCZ data we wished to confirm that neither the QC applied to PGC-SCZ data nor the use of imputed genotypes would introduce artefactual inflation of the estimates of h^2. We found that the adjusted estimate of variance explained for the ISC sample in PGC-SCZ was less in the ISC sample following our QC procedures (analysis 3 vs. analysis 1 and 2). The analyses differ in several ways (imputed v genotyped SNPs, QC on SNPs, QC on samples). In order to exclude

imputation as a source of bias we repeated analysis 3 using only the SNPs that survived PGC-SCZ QC and show that the adjusted estimate of variance explained is the same for genotyped as imputed SNPs (analysis 4 vs. 3). The number of SNPs in analysis 4 is fewer than that from our extreme QC procedure, but this likely reflects less extreme QC on samples. To investigate the impact of sample QC we excluded the ISC London sample from the PGC-SCZ ISC data (analysis 5), as this was the major difference between samples used in our analysis and the PGC analysis. The adjusted h^2 from this analysis is more in line with the estimates from analysis 1 and 2. We conclude that the sample PGC-SCZ QC steps do not introduce artefactual inflation of estimates.

MGS

Our second set of preliminary analyses used the MGS GWAS data set[9] with data on genotyped SNPs accessed with permission from dbGAP. Non-European ancestry outliers based on principal components were excluded. Samples genotyped on plates where the vast majority of samples were either cases or controls were excluded as were samples where the DNA was extracted from cell lines rather than blood, as this DNA source was highly confounded with case-control status. The numbers of SNPs, cases and controls remaining after these QC steps are listed in Supplementary Table 2. We repeated the analyses (1-4) undertaken for the ISC data set using the MGS data set and confirmed no upward biases resulting from using the PGC-SCZ QCed samples nor the imputed genotypes.

Importantly, our analyses demonstrate that the estimates of the proportion of variation in liability to schizophrenia that is tagged by common SNPs are fully consistent between two large independent case-controls studies. The analyses listed in Supplementary Table 2 justified the use of the full PGC-SCZ sample and imputed genotypes as reported in the main paper.

2) Notes on Supplementary Table 3

a) Although the point estimates of the variance explained by X chromosome SNPs is greater in females than in males, the difference is not significant. The variance explained by the X chromosome is consistent with its length (Figure 1). The correlation between the liabilities for schizophrenia based on SNP data from males and females is the same for the X chromosome as for autosomes.
b) The relationship matrix derived from SNPs in the X chromosome was simultaneously fitted with the relationship matrix derived from SNPs in autosomes. The likelihood ratio test statistic to test the null hypothesis that the variance is zero is significant and high for both the equal variance and dosage compensation models and it is difficult to distinguish between them with these data.

3) Notes on Supplementary Table 6

The simulations were conducted to provide insight into the observed results and aimed to establish if very common causal variants could be excluded. First we summarise the observed results:

i) Analysis of the PGC-SCZ data showed that SNPS with MAF > 0.4 explained 5% (s.e. 1%) of the variance in liability (Supplementary Table 5), with approximately equal proportion of variance explained by SNPs in each of the four MAF bins > 0.1.
ii) When genome-wide similarities based only on SNPs with MAF > 0.4 are fitted 11% (s.e. 1%) of the variance in liability is explained (Supplementary Table 5), with approximately equal proportion of variance explained by SNPs in each of the four MAF bins > 0.1.
iii) None of the MAF bins fitted separately can explain the variance explained when all SNPs are used (**Supplementary Table 5**).

Our simulations were designed to explore if a genetic architecture of only rare variants was consistent with our results. Based on the known relationship between allele frequencies and LD, it is highly unlikely that the variance associated with common SNPs reflects only rare causal variants. For example, the maximum LD r^2 possible between variants of frequency 0.4 and 0.01 is only 0.015. In the simulation where 50% of the variance in liability was associated with SNPs with MAF< 0.1 the

joint analysis across MAF bins and the analysis using only individual bins demonstrate that, as expected from the simulation strategy, SNPs with MAF < 0.1 explained close to the true values of variance in liability allocated to them; the results for MAF < 0.1 therefore are verification of the simulation strategy and are not designed for comparison to the observed results. In contrast, under the assumption that the LD architecture is the same between the SNPs with MAF > 0.1 and SNPs in our analysis with MAF < 0.1 as between SNPs with MAF > 0.1 and all causal variants with MAF < 0.1, then the simulation results for MAF > 0.1 can be compared to the observed results.
The analysis using SNPs with MAF > 0.1 captured only ~36% (0.08/0.25, 0.18/0.5, 0.30/0.80) of the true simulated genetic variance. In particular, in the analyses fitting only individual MAF bins the variance explained decreased as the MAF frequency of the SNP bin increased, so that SNPs with MAF > 0.4, only explained ~10% of the total variance associated with the SNPs with MAF < 0.1, compared to 50% (0.11/0.22) in the PGC-SCZ data. The pattern of results across MAF bins in the simulation shows a decreasing proportion of the variance explained with increasing MAF bins whereas the variance attributed to each MAF was approximately equal in the PGC-SCZ analyses.
For comparison, the simulation in which all causal variants have frequency across the allele spectrum the pattern of the proportion of variance explained by each MAF bin is similar to the observed pattern.

The simulation provides a best-case scenario for a rare variants only model, as the causal variation is associated with SNPs <0.1, extending the boundary of rare to uncommon. It is a best-case scenario for all causal variation being rare, since the average r^2 between the SNPs > 0.1 and causal variation will be greater in our simulation than under the usual definition of a rare variants only model.

4) Comparison of our methods and estimates with those of Purcell et al.[4].
Purcell *et al.* used SNP associations detected in the ISC "discovery" sample to construct a linear predictor (a polygenic score) in the MGS "target" sample; the score explained 0.032 of the variance estimated by Nagelkerke's R^2. This R^2 is on the *observed* scale and is relative to a total *variance of case-control status* of the *target* sample ascertained to have approximately an equal number of cases and controls. In contrast, our estimate of h^2 is on the *liability scale* of the *population* (disease prevalence 1%) and is a direct reflection of the properties of the *discovery* sample. However, the critical difference is that R^2 and h^2 are estimates of different parameters and have different properties. For example, to obtain a high R^2, the estimates of effect sizes of individual SNPs need to be accurate, in contrast our estimate of h^2 is based on a single estimate of genome-wide similarity between all pairs of cases and controls and in this way all causal variants can contribute whatever their true effect size as long as they are in LD with the SNPs included in the analysis. Our estimate of h^2 is an unbiased estimator so that the estimate will not systematically increase or decrease with sample size (although the standard error of the estimate will decrease). In contrast, the polygenic score R^2 will increase with the sample size of the discovery sample, because larger sample size should generate more accurate estimates of each individual SNP effects. The PGC-SCZ analysis[1] showed that using ISC+MGS+Cardiff data sets as a discovery sample generated a Nagelkerke's R^2 of 6% in a target sample comprising the other PGC-SCZ data sets. Purcell *et al.* used simulations to calibrate scenarios of variance in liability against the observed R^2 results. These simulations showed that a proportion of variation in liability in the population in LD with genotyped SNPs of 0.34 was consistent with the polygenic score R^2 of 0.032. It is reassuring that our direct estimate of h^2 is consistent with this simulated calibration analysis (Supplementary Table 2 Analysis 1). The ability of the polygenic score derived from the ISC sample to explain variance in the MGS sample reflects that the same SNPs are associated with schizophrenia in two independent samples; we have directly estimated the correlation in liability between the samples tagged by the SNPs as 0.85 (**Table 2**).

5) Comparison of our method with that of Kang et al[10].

Kang *et al*[10] recently presented the efficient mixed model association expedited method (EMMAX) to account for population structure caused by population stratification and hidden relatedness. The method uses a genome-wide identity-by-state (IBS) similarity matrix in an association framework, to correct for a number of possible confounders in the data, including pedigree relationships, cryptic relationships and stratification. Hence the purpose of fitting the similarity matrix is to soak up 'noise' that might affect association signals between individual SNPs and the phenotype because the residuals in the model for association may be correlated. A similar approach was taken recently for a meta-analysis of multiple sclerosis[11]. Kang *et al.* describe the their model in terms of fitting a kinship matrix (i.e., IBD) but the matrix they fit is an IBS (Balding-Nichols) similarity matrix. The similarity matrix fitted by Kang *et al.* estimated from the SNP data is similar to ours and also similar to that used in Eigenstrat[12]. The variance component that is estimated in the Kang *et al.* implementation is an unknown composite of variation due to pedigree relationships (IBD) and due to LD (IBS).

Kang *et al.* called the proportion of total phenotypic variance explained by fitting the IBS matrix 'pseudo-heritability', but the interpretation of their estimate is not obvious. For example, by including relatives in the analysis (Kang *et al*. are likely to have had close relatives in their examples for quantitative traits because the data were from isolated populations in Finland and Sardinia), the SNP matrix estimates approximate identity-by-descent between close relatives (instead of IBS for distant relatives), and therefore the variance explained is driven by the phenotypic correlation of close relatives. Essentially this is like fitting an 'A (additive) E(environmental)' pedigree based model to the data.

In their application of case-control data, Kang *et al*. also use a linear model (like we do) but because their interest was in SNP-disease association, they did not transform the estimate of the variance component to a different scale nor adjust for ascertainment.

In contrast, our approach is to be extremely stringent on SNP and sample QC and estimated relatedness, so that we minimize or avoid possible confounders and are left over with conventionally unrelated individuals whose phenotypes are only correlated due to the proportion of the genome that they share, so that the estimate of variance reflects LD between unknown 'causal' variants and the genotyped SNPs. In Yang *et al.* (2010, NG) we show that our model is statistically equivalent to fitting all SNPs in the model of analysis (as in a GWAS) but with the SNP effects being random and drawn from a distribution of effects.

In summary, operationally the two methods are similar but the rationale, the interpretation of estimated variance components and the treatment of case-control data differ.

REFERENCES

1. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43**, 969-U77 (2011).
2. Altshuler, D.L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
3. Siu, H., Zhu, Y., Jin, L. & Xiong, M. Implication of next-generation sequencing on association studies. *Bmc Genomics* **12**(2011).

4. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752 (2009).
5. Lee, S.H. *et al.* A simple and fast two-locus quality control test to detect false positives due to batch effects in genome-wide association studies. *Genetic Epidemiology* **34**, 854-62 (2010).
6. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-569 (2010).
7. Visscher, P.M., Yang, J. & Goddard, M.E. A commentary on "Common SNPs explain a large proportion of the heritability for human height" by Yang *et al*. (2010). *Twin Research and Human Genetics* **13**, 517-524 (2010).
8. Visscher, P.M., Yang, J. & Goddard, M.E. A Commentary on 'Common SNPs Explain a Large Proportion of the Heritability for Human Height' by Yang et al. (2010). *Twin Research and Human Genetics* **13**, 517-524 (2010).
9. Shi, J.X. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753-757 (2009).
10. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348-U110 (2010).
11. Sawcer, S. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214-219 (2011).
12. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909 (2006).

6) Consortia Authors:

The following authors are included under International Schizophrenia Consortium (ISC) in the Schizophrenia Psychiatric Genome Wide Association Study Consortium (PGC-SCZ):

ISC – Aberdeen: David St Clair (University of Aberdeen, Aberdeen, Scotland). **ISC – Cardiff:** George K. Kirov (Cardiff University, Cardiff, UK), Michael C. O'Donovan (Cardiff University, Cardiff, UK), Peter A. Holmans (Cardiff University, Cardiff, UK), Lyudmila Georgieva (Cardiff University, Cardiff, UK), Ivan Nikolov (Cardiff University, Cardiff, UK), Hywel J. Williams (Cardiff University, Cardiff, UK), Draga Toncheva (University Hospital Maichin Dom, Sofia, Bulgaria), Vihra Milanova (Alexander University Hospital, Sofia, Bulgaria), Michael J. Owen (Cardiff University, Cardiff, UK). **ISC – Dublin:** Derek W. Morris (Trinity College Dublin, Dublin, Ireland), Colm T. O'Dushlaine (Trinity College Dublin, Dublin, Ireland), Elaine Kenny (Trinity College Dublin, Dublin, Ireland), Emma M. Quinn (Trinity College Dublin, Dublin, Ireland), Michael Gill (Trinity College Dublin, Dublin, Ireland), Aiden Corvin (Trinity College Dublin, Dublin, Ireland). **ISC – Edinburgh:** Douglas H. R. Blackwood (University of Edinburgh, Edinburgh, UK), Kevin A. McGhee (University of Edinburgh, Edinburgh, UK), Ben Pickard (University of Strathclyde, Glasgow, UK), Pat Malloy (University of Edinburgh, Edinburgh, UK), Alan W. Maclean (University of Edinburgh, Edinburgh, UK), Andrew McIntosh (University of Edinburgh, Edinburgh, UK). **ISC – London:** Andrew McQuillin (University College London Medical School, London, UK), Khalid Choudhury (University College London Medical School, London, UK), Susmita Datta (University College London Medical School, London, UK), Jonathan Pimm (University College London Medical School, London, UK), Srinivasa Thirumalai (West Berkshire NHS Trust, Reading, UK), Vinay Puri (University College London Medical School, London, UK), Robert Krasucki (University College London Medical School, London, UK), Jacob Lawrence (University College London Medical School, London, UK), Digby Quested (University of Oxford, Oxford, UK), Nicholas Bass (University College London Medical School, London, UK), Hugh Gurling (University College London Medical School, London, UK). **ISC – Portugal:** Michele T. Pato (University of Southern California, Los Angeles, USA), Carlos N. Pato (University of Southern California, Los Angeles, USA), Ayman Fanous (Washington VA Medical Center, Washington, DC, USA, Georgetown University School of Medicine, Washington, DC, USA, and Virginia Commonwealth University School of Medicine, Richmond, USA). **ISC - SW1, ISC - SW2:** Christina M. Hultman (Karolinska Institutet, Stockholm, Sweden), Paul Lichtenstein (Karolinska Institutet, Stockholm, Sweden), Sarah E. Bergen (Massachusetts General Hospital, Boston, USA), Shaun Purcell (Broad Institute, Cambridge, USA), Edward Scolnick (Broad Institute, Cambridge, USA),

Pamela Sklar (Massachusetts General Hospital, Boston, USA, and Mount Sinai School of Medicine, New York, USA), Patrick F. Sullivan (Karolinska Institutet, Stockholm, Sweden, and University of North Carolina, Chapel Hill, USA).

The following authors are included under Molecular Genetics of Schizophrenia Collaboration (MGS) in the Schizophrenia Psychiatric Genome Wide Association Study Consortium (PGC-SCZ):
Pablo V. Gejman (North Shore University Health System, Evanston, USA, and University of Chicago, Chicago, USA), Alan R. Sanders (NorthShore University HealthSystem, Evanston, USA, and University of Chicago, Chicago, USA), Jubao Duan (North Shore University Health System, Evanston, USA, and University of Chicago, Chicago, USA), Douglas F. Levinson (Stanford University, Stanford, USA), Jianxin Shi (National Cancer Institute, Bethesda, USA), Nancy G. Buccola (Louisiana State University, New Orleans, USA), Bryan J. Mowry (Queensland Brain Institute, Brisbane, Australia), Robert Freedman (University of Colorado Denver, Aurora, USA), Farooq Amin (Emory University, Atlanta, USA, and Atlanta Veterans Affairs Medical Center, Atlanta, USA), Donald W. Black (University of Iowa, Iowa City, USA), Jeremy M. Silverman (Mount Sinai School of Medicine, New York, USA, and Veterans Affairs Medical Center, New York, USA), William F. Byerley (University of California at San Francisco, San Francisco, USA, and Northern California Institute for Research And Education, San Francisco, USA), C. Robert Cloninger (Washington University, St. Louis, USA).

The authors are included under other studies (OTH) in the Schizophrenia Psychiatric Genome Wide Association Study Consortium (PGC-SCZ):
SGENE – Bonn: Sven Cichon (University of Bonn, Bonn, Germany, and Research Center Juelich, Juelich, Germany), Marcella Rietschel (University of Bonn, Bonn, Germany, and University of Heidelberg, Mannheim, Germany), Markus M. Nöthen (University of Bonn, Bonn, Germany, and Research Center Juelich, Juelich, Germany), Wolfgang Maier (University of Bonn, Bonn, Germany), Thomas G. Schulze (University of Heidelberg, Mannheim, Germany, and National Institute of Mental Health, Bethesda, USA), Manuel Mattheisen (University of Bonn, Bonn, Germany). **SGENE - Copenhagen**Thomas Hansen (Copenhagen University Hospital, Roskilde, Denmark), Andrés Ingason (Copenhagen University Hospital, Roskilde, Denmark), Henrik B. Rasmussen (Copenhagen University Hospital, Roskilde, Denmark), Line Olsen (Copenhagen University Hospital, Roskilde, Denmark), Henriette Schmock (Copenhagen University Hospital, Roskilde, Denmark), Johan Hilge Thygesen (Copenhagen University Hospital, Roskilde, Denmark), Anders Rosengren (Copenhagen University Hospital, Roskilde, Denmark), Thomas Werge (Copenhagen University Hospital, Roskilde, Denmark). **SGENE - Munich:** Ina Giegling (Ludwig-Maximilians University, Munich, Germany), Annette M. Hartmann (Ludwig-Maximilians University, Munich, Germany), Heike Konnerth (Ludwig-Maximilians University, Munich, Germany), Marion Friedl (Ludwig-Maximilians University, Munich, Germany), Bettina Konte (Ludwig-Maximilians University, Munich, Germany), Pierandrea Muglia (University of Toronto, Toronto, Canada. NeuroSearch A/S, Ballerup, Denmark), Dan Rujescu (Ludwig-Maximilians University, Munich, Germany). **SGENE - TOP3:** Srdjan Djurovic (University of Oslo, Oslo, Norway, and Oslo University Hospital, Oslo, Norway), Morten Mattingsdal (University of Oslo, Oslo, Norway, and Sørlandet Hospital, Kristiansand, Norway), Ingrid Agartz (University of Oslo, Oslo, Norway, and Diakonhjemmet Hospital, Oslo, Norway), Ingrid Melle (University of Oslo, Oslo, Norway, and Oslo University Hospital, Oslo, Norway), Ole A. Andreassen (University of Oslo, Oslo, Norway, and Oslo University Hospital, Oslo, Norway). **SGENE – UCLA:** Roel A. Ophoff (University Medical Center Utrecht, Utrecht, The Netherlands, and University of California at Los Angeles, Los Angeles, USA), Rita M. Cantor (University of California at Los Angeles, Los Angeles, USA), Nelson B. Freimer (University of California at Los Angeles, Los Angeles, USA), René S. Kahn (University Medical Center Utrecht, Utrecht, The Netherlands), Don H. Linszen (University of Amsterdam, Amsterdam, The Netherlands), Jim van Os (Maastricht University Medical Centre, Maastricht, The Netherlands), Durk Wiersma (University of Groningen, Groningen, The Netherlands), Richard Bruggeman (University of Groningen, Groningen, The Netherlands), Wiepke Cahn (University Medical Center Utrecht, Utrecht,

The Netherlands), Lieuwe de Haan (Academic Medical Centre University of Amsterdam, Amsterdam, The Netherlands), Lydia Krabbendam (Maastricht University Medical Centre, Maastricht, The Netherlands), Inez Myin-Germeys (Maastricht University Medical Centre, Maastricht, The Netherlands), Eric Strengman (University Medical Center Utrecht, Utrecht, The Netherlands). **Zucker Hillside:** Anil K. Malhotra (The Zucker Hillside Hospital Division of the North Shore, Glen Oaks, USA, The Feinstein Institute for Medical Research, Manhasset, USA, and Albert Einstein College of Medicine of Yeshiva University, Bronx, New York, USA), Todd Lencz (The Zucker Hillside Hospital Division of the North Shore, Glen Oaks, USA, The Feinstein Institute for Medical Research, Manhasset, USA, and Albert Einstein College of Medicine of Yeshiva University, Bronx, New York, USA). **Cardiff UK:** Michael C. O'Donovan (Cardiff University, Cardiff, UK), Nicholas Craddock (Cardiff University, Cardiff, UK), Peter A. Holmans (Cardiff University, Cardiff, UK), Marian Hamshere (Cardiff University, Cardiff, UK), Hywel J. Williams (Cardiff University, Cardiff, UK), Valentina Moskvina (Cardiff University, Cardiff, UK), Sarah Dwyer (Cardiff University, Cardiff, UK), Lyudmila Georgieva (Cardiff University, Cardiff, UK), Stan Zammit (Cardiff University, Cardiff, UK), Michael J. Owen (Cardiff University, Cardiff, UK). **CATIE**: Patrick F. Sullivan (Karolinska Institutet, Stockholm, Sweden, and University of North Carolina, Chapel Hill, USA), Dan-Yu Lin (University of North Carolina, Chapel Hill, USA), Edwin van den Oord (Virginia Commonwealth University, Richmond, USA), Yunjung Kim (University of North Carolina at Chapel Hill, Chapel Hill, USA), T. Scott Stroup (Columbia University, New York, USA), Jeffrey A Lieberman (Columbia University, New York, USA). **PGC-SCZ additional:** Kenneth S. Kendler (Virginia Commonwealth University, Richmond, USA), Mark J Daly (Center for Human Genetic Research, Massachusetts General Hospital, Boston, USA), D Posthuma (Faculty of Psychology and Education of VU University, Amsterdam).

7) PGC-SCZ Acknowledgements

PGC-SCZ Overall Coordination: Dr. Gejman's efforts were supported by NIH grants (R01 MH59571 to P.V.G.; R01 MH81800 to P.V.G.; U01 MH79469 to P.V.G.; and U01 MH85508 to P.V.G.), and by The Paul Michael Donovan Charitable Foundation. Dr. Daly's efforts were supported by NIMH U01 MH85515. Dr. Kendler's efforts were supported by R01 MH83074. Analytical activities of the PGC were supported NIMH U01 grants (MH85520, MH85518, MH85515, MH85513, and MH85508 with PIs Sullivan, Faraone, Daly, Purcell, and Gejman). Additional analytical support was from Foundation for the NIH (grant ID BROAD09GAIN0 – PI Daly) and R01 MH80403 (PI Sullivan). All computational work was conducted on the Genetic Cluster Computer (the Netherlands) which is funded by an NOW Medium Investment grant (480-05-003, PI Posthuma), the Faculty of Psychology and Education of VU University (Amsterdam), and by the Dutch Brain Foundation (PI Ophoff) and is hosted by the Dutch National Computing and Networking Services. Dr. Lin's efforts were supported by NIH grants R37 GM47845, R01 CA82659, and P01 CA142538. Dr. Posthuma is financially supported by the Netherlands Organisation for Scientific Research (NWO 016-065-318; 40-00812-98-07-032) and the Neuroscience Campus Amsterdam. Dr. Gejman's and Sanders' efforts were supported by NIH grants (R01 MH59571 to P.V.G.; R01 MH81800 to P.V.G.; U01 MH79469 to P.V.G.; and U01 MH85508 to P.V.G.) and by The Paul Michael Donovan Charitable Foundation. Dr. Fanous is or has been supported by grants from the Department of Veterans Affairs Merit Review Program. Dr. Kendler's efforts were supported by R01 MH83074.

Individual study samples (Stage 1 of the PGC-SCZ analysis[1]).

Stage 1: GWAS – European ancestry sample 1 – Cardiff UK. The Cardiff Group members are supported by grants from the MRC, the Wellcome Trust and by a NIMH (USA) CONTE: 2 P50 MH066392-05A1. This study makes use of control data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. We would also like to acknowledge J. L. Marchini, C. Spencer, B. Howie, and H-T. Leung who were involved in making the genotype calls in this dataset for the primary manuscript.

Stage 1: GWAS – European ancestry sample 2 – CATIE. Dr. Sullivan was supported by R01s

MH074027 and MH077139. The CATIE project was funded by NIMH contract N01 MH90001. Control subjects from the National Institute of Mental Health Schizophrenia Genetics Initiative (NIMH-GI), data and biomaterials were collected by the "Molecular Genetics of Schizophrenia II" (MGS-2) collaboration. The investigators and coinvestigators were: NorthShore University HealthSystem, Evanston, IL, R01 MH59571, Pablo V. Gejman, M.D. (Collaboration Coordinator; PI), Alan R. Sanders, M.D.; Emory University School of Medicine, Atlanta, GA, R01 MH59587, Farooq Amin, M.D. (PI); Louisiana State University Health Sciences Center, New Orleans, LA, R01 MH67257, Nancy G. Buccola APRN, B.C., M.S.N. (PI); University of California-San Francisco, San Francisco, CA, R01 MH60870, William F. Byerley, M.D. (PI); Washington University, St Louis, MO, U01, MH60879, C. Robert Cloninger, M.D. (PI); University of Iowa, Iowa, IA, R01 MH59566, Donald W. Black, M.D. (PI), Raymond R. Crowe, M.D.; University of Colorado, Denver, CO, R01 MH59565, Robert Freedman, M.D. (PI); Stanford University, Palo Alto, CA, R01 MH61675, Douglas F. Levinson MD (PI); University of Queensland, Brisbane, Queensland, Australia; R01 MH59588, Bryan J. Mowry, MD (PI); Mt Sinai School of Medicine, New York, NY, R01 MH59586, Jeremy M. Silverman, Ph.D. (PI).
Stage 1: GWAS – European ancestry sample 3 – ISC – Aberdeen. The work at the University of Aberdeen was partly funded by GlaxoSmithKline and Generation Scotland, Genetics Health Initiative.
Stage 1: GWAS – European ancestry sample 4 – ISC – Cardiff. The Cardiff University group was supported by a Medical Research Council (UK) Programme grant and the National Institutes of Mental Health (USA) (CONTE: 2 P50 MH066392-05A1).
Stage 1: GWAS – European ancestry sample 5 – ISC – Dublin. The Trinity College Dublin group was supported by Science Foundation Ireland, the Health Research Board (Ireland), the Stanley Medical Research Institute and the Wellcome Trust; Irish controls were supplied by J. McPartlin from the Trinity College Biobank.
Stage 1: GWAS – European ancestry sample 6 – ISC – Edinburgh. The collection of the University of Edinburgh cohort was supported by the Wellcome Trust Clinical Research Facility (Edinburgh) and grants from The Wellcome Trust, London and the Chief Scientist Office of the Scottish Government. B. Pickard held a Sim Fellowship from the Royal College of Physicians in Edinburgh. We acknowledge the help of M. Van Beck in gathering patient samples and data and L. Murphy for DNA preparation and sample archiving at the Wellcome Trust Clinical Research Facility, Edinburgh.
Stage 1: GWAS – European ancestry sample 7 – ISC – London. University College London clinical and control samples were collected with support from the Neuroscience Research Charitable Trust, the Camden and Islington Mental Health and Social Care Trust, East London and City Mental Heath Trust, the West Berkshire NHS Trust, the West London Mental Health Trust, Oxfordshire and Buckinghamshire Mental Health Partnership NHS Trust, South Essex Partnership NHS Foundation Trust, Gloucestershire Partnership NHS Foundation Trust, Mersey Care NHS Trust, Hampshire Partnership NHS Trust and the North East London Mental Health Trust.
Stage 1: GWAS – European ancestry sample 8 – ISC – Portugal. CNP and MTP are or have been supported by grants from the NIMH (MH085548, MH085542, MH071681, MH061884, MH58693, and MH52618) and the NCRR (RR026075). CNP, MTP, and AHF are or have been supported by grants from the Department of Veterans Affairs Merit Review Program.
Combined acknowledgements for: Stage 1: GWAS – European ancestry sample 9 – ISC - SW1; Stage 1: GWAS – European ancestry sample 10 – ISC – SW2. The group at the Karolinska Institutet was supported by the Swedish Council for Working Life and Social Research (FO 184/2000; 2001-2368). The group at the University of North Carolina, Chapel Hill, was supported by MH074027,MH077139, MH080403, and MH085520, the Sylvan C. Herman Foundation (P.F.S.) and the Stanley Medical Research Institute (P.F.S.).
Stage 1: GWAS – European ancestry sample 11 – MGS. We thank the study participants, and the research staff at the study sites. This study was supported by NIH R01 grants (MH67257 to N.G.B., MH59588 to B.J.M., MH59571 to P.V.G., MH59565 to R.F., MH59587 to F.A., MH60870 to W.F.B., MH59566 to D.W.B., MH59586 to J.M.S., MH61675 to D.F.L., MH60879 to C.R.C., and MH81800 to P.V.G.), NIH U01 grants (MH46276 to C.R.C., MH46289 to C. Kaufmann, MH46318 to M.T. Tsuang,

MH79469 to P.V.G., and MH79470 to D.F.L.), the Genetic Association Information Network (GAIN), and by The Paul Michael Donovan Charitable Foundation. Genotyping was carried out by the Center for Genotyping and Analysis at the Broad Institute of Harvard and MIT (S. Gabriel and D. B. Mirel), which is supported by grant U54 RR020278 from the National Center for Research Resources. Genotyping of half of the EA sample and almost all the AA sample was carried out with support from GAIN. The GAIN quality control team (G.R. Abecasis and J. Paschall) made important contributions to the project. We thank S. Purcell for assistance with PLINK.

Stage 1: GWAS – European ancestry sample 12 – SGENE – Bonn. This study was supported by the German Federal Ministry of Education and Research (BMBF), within the context of the National Genome Research Network 2 (NGFN-2), the National Genome Research Network plus (NGFNplus), and the Integrated Genome Research Network (IG) MooDS (grant 01GS08144 to S.C. and M.M.N., grant 01GS08147 to M.R.). M.M.N. also received support from the Alfried Krupp von Bohlen und Halbach-Stiftung. We are grateful to K.- H. Jöckel and R. Erbel for providing control individuals from the Heinz Nixdorf Recall Study, to S. Schreiber for providing control individuals from the PopGen study, and to H.-E. Wichmann for providing control individuals from the KORA study.

Stage 1: GWAS – European ancestry sample 13– SGENE – Copenhagen. The study was sponsored by grant to TW from the Lundbeck Foundation (No R34-A3243), the Danish National Advanced Technology Foundation (No 001-2009-2), the Danish Medical Research Council (No 09-065634), the European Union Marie Curie Program (Project PsychGene; No PIAP-GA-2008-218251), and the Danish Psychiatric Research Foundation.

Stage 1: GWAS – European ancestry sample 14– SGENE – Munich
We thank David Goldstein and colleagues for genotyping parts of the GWAS sample from Munich.

Stage 1: GWAS – European ancestry sample 15 – SGENE - TOP3. We thank the TOP study group members for their contribution to data collection. The work was supported by grants from the Research Council of Norway (#167153/V50, #163070/V50, #175345/V50); South-East Norway Health Authority (#123-2004); Oslo University Hospital and University of Oslo. E. Lilly Inc supported parts of the genotyping costs.

Stage 1: GWAS – European ancestry sample 16 – SGENE – UCLA. We thank Harry van Someren for database management. Funding was provided by R01 MH078075 (R.A.O.).

Stage 1: GWAS – European ancestry sample 17 – Zucker Hillside. The ZHH GWAS was supported by the Donald and Barbara Zucker Foundation, internal funding from the North Shore – Long Island Jewish Health System, and grants from National Alliance for Research on Schizophrenia and Depression (to AKM), and the National Institutes of Health (K23MH001760 and R01MH079800 to AKM; R01MH0084098 to TL; and Center grants P30MH074543 to John M. Kane and M01RR018535 to Kevin J. Tracey).